Authorship Attribution: Specifics for Slovene

Ana Zwitter Vitez

Trojina, Zavod za uporabno slovenistiko, Cesta v Kleče 16, SI – 1000 Ljubljana, ana.zwitter@guest.arnes.si

SCN IV/2 [2011], 75–85

V prispevku je poudarjena pomembnost kakovostne analize jezikovnih parametrov, ki omogočajo ugotavljanje avtorstva ali profiliranje avtorja besedila v forenzičnem, literarnozgodovinskem ali gospodarskem kontekstu (anonimna grozilna pisma, ugotavljanje plagiatorstva, literarna besedila neznanega izvora, profiliranje strank). Ker je tovrstne analize za slovenščino težko najti, predlagamo metodologijo luščenja skladenjskih, leksikalnih, semantičnih in znakovnih parametrov za potrebe kvantitativne obravnave avtorjevega osebnega sloga.

The paper shows the importance of a quality analysis of linguistic features which enable the process of authorship attribution or author profiling in a forensic, literary or economic context (anonymous threat letters, plagiarism, literary works of unknown authorship, client profiling). It also highlights the lack of realized analyses for Slovene and outlines the methodology of detecting the syntactic, lexical, semantic and character features in order to quantify the author's personal style.

Ključne besede: ugotavljanje avtorstva besedil, profiliranje avtorja, jezikovni parametri, jezikovne tehnologije, forenzično jezikoslovje

Key words: authorship attribution, author profiling, linguistic features, language technologies, forensic linguistics

1 Introduction

Authorship attribution has been a hot topic of interest in criminology and literary history since the end of the 19th century. A pioneering study in authorship attribution using the linguistic features of texts was conducted by Thomas Corwin Mendenhall (1901). He analysed the length of words to identify the differences between different languages and different authors, and discovered that Shakespeare's and Marlowe's histograms are almost identical (Marlowe's

death two weeks before the first publication of Shakespeare's first works is still unexplained).

Another famous example represents the Mosteller's and Wallace's study which answers the question who wrote the disputed Federalist papers (promoting the ratification of the United States Constitution) by comparing different statistical methods.

Nowadays, the availability of language corpora, machine learning methods and data mining enable the further development of methods for authorship attribution. It is therefore not surprising that international research on authorship attribution is no longer limited to attributing the authorship of a text to one of the potential authors (Stamatatos et al. 2001), but has also developed subfields such as plagiarism detection (Meyer zu Eissen et al. 2007), author profiling (Koppel et al. 2002) and detection of stylistic inconsistencies in collaborative writing (Graham et al. 2005).

2 State of the art

One of the key aims of our research into authorship attribution is to identify style markers (i.e., linguistic features) that quantitatively specify the author's style (Rudman 1998). Next, we offer an overview of the latest research according to four types of linguistic features: lexical features, character features, syntactic features and semantic features.

Lexical features

Lexical features of a text are normally presented using word frequency vectors (Sebastiani 2002). The most common words (articles, prepositions, pronouns, etc.) are found to be among the best lexical features to discriminate between authors because they are used unconsciously by the authors (Burrows 1987; Argamon and Levitan 2005; Luyckx and Daelemans 2005). However, Stamatatos (2009) and Eder (2010) point out that the success of the lexical features method is largely dependent on the length of the text.

Character features

Character features of texts (letter and punctuation frequency) are also very useful in quantifying the author's style (Grieve 2007). A more elaborate approach is to extract frequencies of n-grams on the character level. Studies by Peng et al. (2003), Keselj et al. (2003), Stamatatos (2006) and Diederich et al. (2003) have yielded good results using n-grams to quantify author's style. A comparative study of lexical and character features of the same corpus (Grieve 2007) showed that n-grams are the most effective measures of authorial style.

Syntactic features

The method of measuring syntactic features in texts is based on the idea that authors tend to use similar syntactic patterns unconsciously. Therefore, syntactic information is considered more reliable than lexical information when determining the author's style. However, syntactic information also requires more advanced tools for natural language processing (e.g., POS tagger, parser). The syntactic features method was first used by Baayen et al. (1996). Since then, the method has been used in several studies, e.g., Stamatatos et al. (2000; 2001), Luyckx and Daelemans (2005), Uzuner and Katz (2005), and Hirst and Feiguina (2007).

Semantic features

Semantic feature extraction from a text is based on the WordNet semantic network, which enables searches for synonyms and hypernyms of words. WordNet has been used in several studies, one of the best known being McCarthy et al. (2006), where the authors also attempt to detect semantic similarities between words by applying latent semantic analysis by Deerwester et al. (1990) to lexical features.

The corpora used in authorship studies are almost always genre-specific, so that authorship is the most important discriminatory factor between the texts (Stamatatos 2009). Stamatatos, however, suggests that any attribution method should be tested on texts with at least one feature (e.g., genre, length or the number of candidate authors) that is different from those used in the training corpus, in order to determine its efficiency and limitations.

All of the studies mentioned use corpus data that include texts annotated with information about authors. Studies dealing with author profiling require more detailed author information; besides name, gender and age, the studies include information on education level, region and the author's psychological profile (Luyckx and Daelmans 2005). Such corpora with author information represent a valuable national document and can be used for further research into authorship attribution and author profiling.

3 Authorship studies for Slovene

In Slovenia, there are only two studies using statistical methods for the purposes of authorship attribution, one using word and sentence length to detect plagiarism (Dović 2002), and the other analysing function words as the potential linguistic feature for authorship attribution (Limbek 2008). This gap in research is preventing more extensive use of authorship attribution and author profiling in authorship law, criminology, literary studies and market research.

The field of authorship attribution is closely connected with the availability of language resources (corpora) and tools (taggers, parsers). The good news

in this area is that we now have language resources and tools for Slovene that demonstrate the following potential:

Tool or resource for Slovene	Potential
The billion-word corpus Gigafida, developed as part of the project Communication in Slovene (http://demo.gigafida.net).	Source of texts for a corpus that can be used for the purposes of authorship attribution and author profiling.
The collection of electronic texts **Slovenska leposlovna klasika** – Slovene classic litterary works (http://sl.wikisource.org/wiki/Glavna_stran).	Slovene litterary works in more or less uniformed electronic form.
Part-of-speech tagger (http://oznacevalnik.slovenscina.eu) and **Parser** (http://razclenjevalnik.slovenscina.eu/), developed as part of the Communication in Slovene project.	Basis for statistical analysis and identification of the lexical, character, syntactic and semantic features decisive for a quantitative description of an author's writing style.

The abovementioned facts clearly show that the important topic of authorship attribution is still under-researched in Slovenia, but that there is good potential for quality research due to the availability of language tools and resources for Slovene. This is why a quality authorship study identifying the linguistic features that reveal the author's personal profile would make a considerable contribution to the progress of criminology, literary studies and market research.

4 What can be done

Research in authorship attribution can answer questions concerning:
- which of the potential authors is the author of a text of unknown authorship, and
- the profile of the author (gender, age, level of education, region, psychological profile) of a text of unknown authorship when no potential authors are available.

The main hypothesis is based on the fact that by using a well designed and annotated corpus of texts we can distinguish linguistic features for determining the author's style in Slovene texts. By identifying sets of linguistic features we can attribute the authorship of a text of an unknown author to one of the potential authors, or, when no potential authors are available, describe the profile of the author (gender, age, level of education, region, psychometric traits).

This knowledge can be gained in the following way:
- building a reference corpus,
- determination and evaluation of lexical, character, syntactic and semantic features for authorship attribution,
- design and evaluation of feature-based models for author attribution and author profiling.

5 The proposed methodology

The scientific approach of the proposed research combines existing language resources and tools for Slovene with knowledge from the fields of corpus linguistics and statistical data analysis, in order to enable the identification of linguistic features that can be used for quantifying the author's writing style. These identified features can determine whether a text of unknown authorship has been written by one of the potential authors, or they can establish the profile of the author where no potential authors are available.

In order to identify the linguistic features that determine the author's personal profile for Slovene it seems reasonable to use the following methodology:

STAGE 1: Design and creation of a reference corpus

a) Specifications for text selection.
- gathering the texts from various sources, including existing corpora of Slovene, websites and individuals
- collection of metatextual information, such as the genre and year of the text, as well as the gender, age and level of education of the author(s).

b) Psychometric data about the authors.
- use of the International Personality Item Pool (IPIP) questionnaire (the 300-question version has been translated into Slovene by Dr Janek Musek) in order to obtain the psychometric traits of agreeableness, extraversion, neuroticism, conscientiousness and openness (the Big Five).
- distribution of the questionnaire to a selection of authors of the collected texts
- calculation of scale scores for each of the five traits.

c) Collection of texts and data preparation.
- validation, normalisation, cleaning and annotation of the collected texts (with metatextual information, including scale scores for psychometric traits)
- tagging and parsing of the collected texts
- compilation of training corpora and test corpora for authorship attribution, as well as a reference corpus for author profiling
- selection of texts and subcorpora for the evaluation stages.

STAGE 2: Design of the authorship attribution models

This stage will identify the best linguistic feature, or combination of linguistic features, for authorship attribution in Slovene, as well as the most appropriate statistical method. Analyses should be performed on corpora containing texts from different genres (for example, newspaper texts, literary texts and texts with a clear conative function).

d) Authorship attribution method for newspaper texts
- measuring of different linguistic features, from lexical features (e.g., vocabulary richness, word frequencies) and character features (e.g., character n-grams), to syntactic features (e.g., chunks) and semantic features (e.g., synonyms)
- use of different statistical methods, such as Naive Bayes, support vector machine (SVM), etc.
- checking of the results with manual language analysis
- formation of subgroups or lists of relevant linguistic parameters.

The result will be the identification of the best (combination of) linguistic features that can be used in authorship attribution for newspaper texts.

e) Authorship attribution method for literary texts
- repetition of the procedure from d) on a corpus of literary texts.
- identification of the combination of linguistic features that produces the best results in attributing authorship to literary texts
- comparison of linguistic features for literary texts with the linguistic features for newspaper texts

f) Authorship attribution method for texts with a clear conative function
- repetition of the procedures described in d) and e) on a corpus of texts with a clear conative function (Jakobson 1960). Examples of such texts are letters from readers, threat letters, etc.
- identification of the linguistic parameters for authorship attribution in conative texts
- comparison of linguistic features for literary and newspaper texts with the linguistic features for conative texts.

STAGE 3: Identification of the best linguistic features for author profiling

In this stage, the reference corpus will be statistically analysed to extract the distinguishing classifiers for different profile categories; namely, gender, age, region (geographic origin), education level and psychometric traits. The problem addressed here is determining the profile of an author of an unknown text when no candidate authors, or their texts, are available.

g) Linguistic features for author profiling in newspaper texts
- determination of distinguishing linguistic features for each author characteristic (gender, age, level of education, region, psychometric traits) in the corpus of newspaper texts.
- grouping of the different features into feature groups (e.g., lexical, character).

h) Linguistic features for author profiling in literary texts
- repetition of procedure g) on a corpus of literary texts
- identifying the optimal linguistic features for author profiling on literary texts

– comparison of the results to that of g).

i) Linguistic features for author profiling in texts with a clear conative function
   – repetition of procedure g) on a corpus of conative texts
   – identifying the optimal linguistic features for author profiling on conative texts (threat letters)
   – comparison of the results to that of g).

STAGE 4: Evaluation

j) Evaluation of the models for authorship attribution
   – determining the success rate of the model to correctly attribute the authorship of an unknown text.

   To evaluate the model, the best method is to use several corpora that differ from the training corpora used in Stage 2 in one characteristic; for example, in the number of texts per author, in the number of candidate authors and in the different length of the texts. In this way, we can evaluate how successful the models are in attributing authorship to unknown texts when at least one variable is different to the one used in the model design.

k) Evaluation of the models for author profiling
   – determining the success rate of the model in determining the author's profile, based on characteristics such as gender, age, region, level of education and psychometric traits.

   It seems wise to use texts from the same genres used in Stage 3, but with different characteristics (e.g., different length).

6 The results and the possible applications of authorship attribution

The results of the proposed research in authorship attribution will be as follows:
– a reference database containing texts that include information on authors (gender, age, level of education, region and psychometric traits),
– a description of methods and optimal linguistic features for authorship attribution in different genres,
– a description of methods and optimal linguistic features for author profiling (determining gender, age, level of education, region and psychological traits) in different genres.

7 The possible applications of authorship attribution

The need for more research on an author's distinguishing linguistic features in a text can be justified by the fact that in Slovenia (and internationally) public figures and individuals are increasingly exposed to threat letters in traditional

or internet form. In the last few years, examples of such public figures include J. Janša, K. Kresal, R. Žerjav, Z. Jelinčič, B. Magajna and R. Batelli.

Due to the accessibility of texts on the web, plagiarism represents a serious problem of intellectual property. This phenomenon was clearly exposed in recent Slovenian political scandals concerning J. Janša's 15th anniversary independence speech, M. Cvikl's BA dissertation and the parliament representative B. Marinič's German test.

In literary studies, a serious study could have solved authorship problems such as those related to the Slovene poem *Oj Triglav, moj dom* (1894), the drama *Ekshibicionist* (2001), written under the pseudonym O. J. Traven, the pornographic novel Čudoviti Klon (2006), published under the pseudonym Eva Pacher, and others.

The recruitment process (human resources) often regards people as the capital and potential of companies (cf. Schuler R. E., Jackson S. E. eds. [1999]): "The conversation leads to the matter of choosing *the right people*. What is the right profile for a certain company? How to recognise it?" The availability of linguistic features, decisive for establishing an author's personal profile, can also contribute towards choosing the right candidate for the job.

Finally, in all fields of the economy knowing the buyer's profile is very important nowadays. For this reason, companies are building databases of clients with different shopping habits and basing their strategies for satisfying customers on them (Shaw et al. 2001). With a methodology allowing client profiling on the basis of linguistic features it will be possible to provide different companies with an opportunity to enhance their databases with the linguistic profiles of clients.

8 Conclusion

The paper shows the importance of a quality analysis of linguistic features in order to enable the quantification an author's personal style. It also highlights the lack of realised analyses for Slovene, as well as outlining the methodology and the possible applications of the proposed work.

Two scenarios seem to be probable: the research will be funded either publicly or commercially. If the first scenario is realised, authorship attribution studies for Slovene will be developed for the purposes of detecting the authors of anonymous threats, plagiarism and literary works, whereas if such research is funded by private companies the studies will mostly contribute to the domain of determining the client's profile. Ethical standards certainly suggest that it is preferable for public agencies to take the lead in such research, but the state's administration will have to make the final decision in this regard[1].

[1] The first Slovenian project concerning authorship attribution and author profiling has been proposed by Trojina, Institute for applied Slovene studies. At the current moment, we are waiting for the results of the Slovenian Research Agency's annual call for proposals.

REFERENCES

Web pages

Corpus Gigafida (http://demo.gigafida.net, 25.05.11)

Slovenska leposlovna klasika (http://sl.wikisource.org/wiki/Glavna_stran, 25.5.11).

Slovenian part-of-speech tagger (http://oznacevalnik.slovenscina.eu, 25.5.11)

Slovenian parser (http://razclenjevalnik.slovenscina.eu/, 25.5.11)

Publications

Shlomo ARGAMON, Shlomo LEVITAN, 2005: Measuring the usefulness of function words for authorship attribution. *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.*

Arald BAAYEN, Hans VAN HALTEREN, Fiona TWEEDIE, 1996: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11/3, 121–131.

John F. BURROWS, 1987: *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Scott DEERWESTER et al., 1990: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407.

Joachim DIEDERICH et al., 2003: Authorship attribution with support vector machines. *Applied Intelligence* 19/1–2, 109–123.

Marijan DOVIĆ, 2002: Podbevšek in Cvelbar: Poskus empirične preverbe namigov o plagiatorstvu. *Slavistična revija* 50, 233–249.

Maciej EDER, 2010: *Does Size Matter? Authorship Attribution, Small Samples, Big Problem.* London: Proceedings of the Digital Humanities Conference 2010.

Neil GRAHAM et al., 2005: Segmenting documents by stylistic character. *Journal of Natural Language Engineering*, 11(4), 397–415.

Jack GRIEVE, 2007: Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270.

Graeme HIRST, Ol'ga FEIGUINA, 2007: Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22/4, 405–417.

Moshe KOPPEL et al., 2002: Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Marko LIMBEK, 2008: Usage of Multivariate Analysis in Authorship Attribution: Did Janez Mencinger Write the Story "Poštena Bohinčeka"? *Metodološki zvezki*, 5/1, 81–93.

Kim LUYCKX, Walter DAELEMANS, 2005: Shallow text analysis and machine learning for authorship attribution. *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands.*

Philip MCCARTHY et al. 2006: Analyzing writing styles with coh-metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference*, 764–769.

Sven MEYER ZU EISSEN et al., 2007: Plagiarism detection without reference collections. *Advances in Data Analysis*, 359–366.

Frederick MOSTELLER, David L. WALLACE, 1964: Inference and Disputed Authorship: the Federalist Papers, Reading, Mass.: Addison-Wesley.

Fuchun PENG et al., 2003: Language independent authorship attribution using character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 267–274.

Joseph RUDMAN, 1998: The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351–365.

Fabrizio SEBASTIANI, 2002: Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1).

Michael SHAW et al. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, 31 (1), 127–137.

Efstathios STAMATATOS et al. 2000: Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4), 471–495.

Efstathios STAMATATOS et al., 2001: Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2), 193–214.

Efstathios STAMATATOS et al., 2006: Ensemble-based author identification using character n-grams. *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, 41–46.

Efstathios STAMATATOS 2009: A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Ozlem UZUNER, Boris KATZ, 2005: A comparative study of language models for book and author recognition. *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, 969–980.

UGOTAVLJANJE AVTORSTVA BESEDIL ZA SLOVENŠČINO

Področje ugotavljanja avtorstva besedil v zadnjih dveh desetletjih doživlja silovit razmah, saj se javne in nejavne osebnosti pogosto srečujejo s pojavom internetnih groženj in grozilnih pisem v tradicionalni obliki**,** poleg tega pa je zaradi lahke dostopnosti besedil na spletu vse bolj prisoten pojav plagiatorstva. Kljub izredno razvitim študijam v mednarodnem merilu to pomembno področje za v Sloveniji ostaja precej neraziskano, vendar obstajajo dobre možnosti za kakovostne raziskave zaradi dobro razvitih jezikovnih orodij in virov za slovenščino.

Ugotavljanje avtorstva besedil temelji na odkrivanju tistih jezikovnih parametrov, na podlagi katerih lahko besedilo neznanega izvora pripišemo določenemu avtorju ali eni od lastnosti avtorjevega profila (spol, starost, izobrazba, regija). Te jezikovne parametre

lahko izluščimo z naslednjo metodologijo: (1) izdelava referenčne baze označenih besedil, (2) ugotavljanje leksikalnih, znakovnih, skladenjskih in semantičnih lastnosti posameznih kategorij za ugotavljanje avtorstva besedila in profila avtorja, (3) izdelava in evalvacija modela za ugotavljanje osebnega profila avtorja.

Končni rezultat take raziskave so izluščeni jezikovni parametri za slovenščino, na podlagi katerih je mogoče ugotoviti, kateri od potencialnih avtorjev je tvoril besedilo neznanega izvora, ali določiti osebni profil neznanega avtorja besedila (spol, starost, izobrazbo, regionalno pripadnost in psihometrične lastnosti). Rezultati raziskave lahko znatno izboljšajo kakovost kriminalističnega preiskovanja, prava avtorskih pravic, proučevanja literarne zgodovine in profiliranja strank za potrebe tržnih analiz.